



PROYECTO FINAL

MYSTERYLAND

OBJETIVOS

OBJETIVO GENERAL

» Realizar la recolección y el análisis de una gran cantidad de datos acerca del evento musical Mysteryland, a realizarse en el mes de Agosto, recopilados a través de tweets.

OBJETIVOS ESPECÍFICOS

- ❖ Establecer funciones del equipo de trabajo.
- ❖ Recopilar tweets acerca del evento musical Mysteryland.
- ❖ Realizar la indexación a través de los nodos.
- ❖ Realizar consultas Map Reduce.
- ❖ Filtrar y analizar los datos.
- ❖ Concluir los resultados analizados.

DEFINICIÓN DEL CASO DE ESTUDIO

El proyecto a realizarse se enfoca en el análisis de datos obtenidos, mediante la red social Twitter durante el transcurso de la realización del festival Mysteryland el 26 y 27 de Agosto del 2017. En el almacén de datos unificado se almacenará los datos recolectados de los diferentes países los datos tendrá diferentes origen (diferentes ordenadores), por el cual la replicación de las bases será necesario para poder obtener una base unificada que nos permita generar informes.

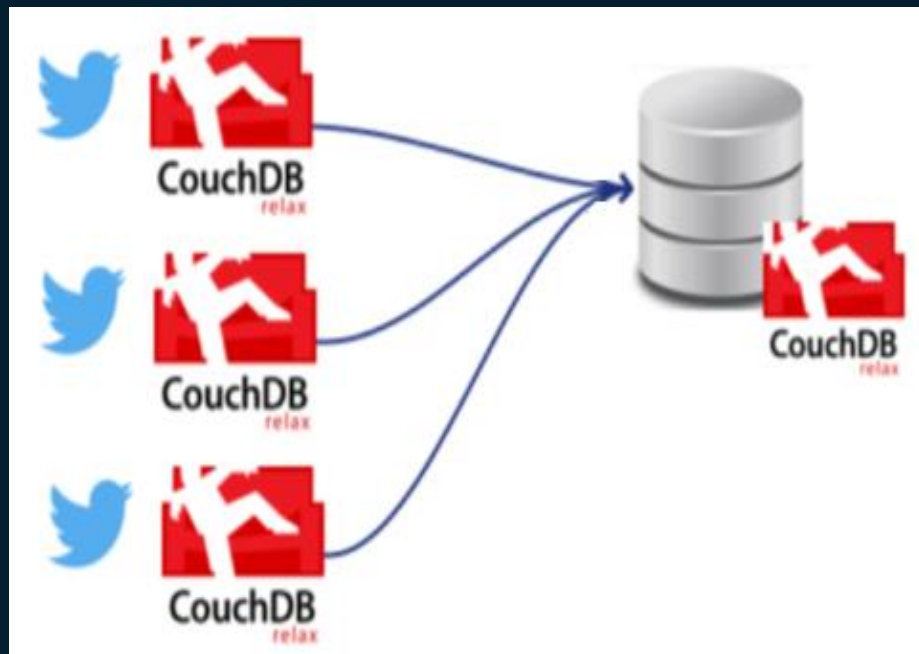
FUNDAMENTO TEÓRICO

En la actualidad el uso de las redes sociales como: Twitter, LinkedIn, Instagram, Facebook, ha permitido que los individuos puedan compartir interés o actividades. El presente proyecto utilizó la red social Twitter, para poder recoger tweets con el #Mysteryland, contando al momento con 224379 de tweets.

RECURSOS Y HERRAMIENTAS

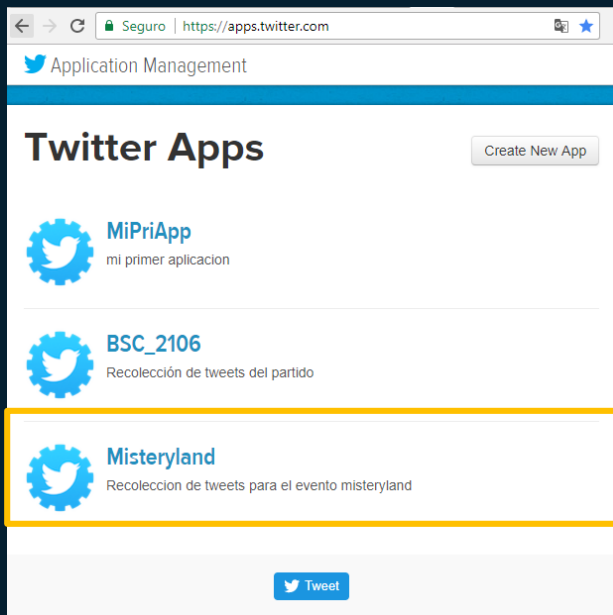
- ❖ Elasticsearch
- ❖ CouchDB
- ❖ Cerebro
- ❖ Bouding Box
- ❖ Kibana
- ❖ Logstash
- ❖ Curl
- ❖ App Twitter
- ❖ Sublime text

ARQUITECTURA

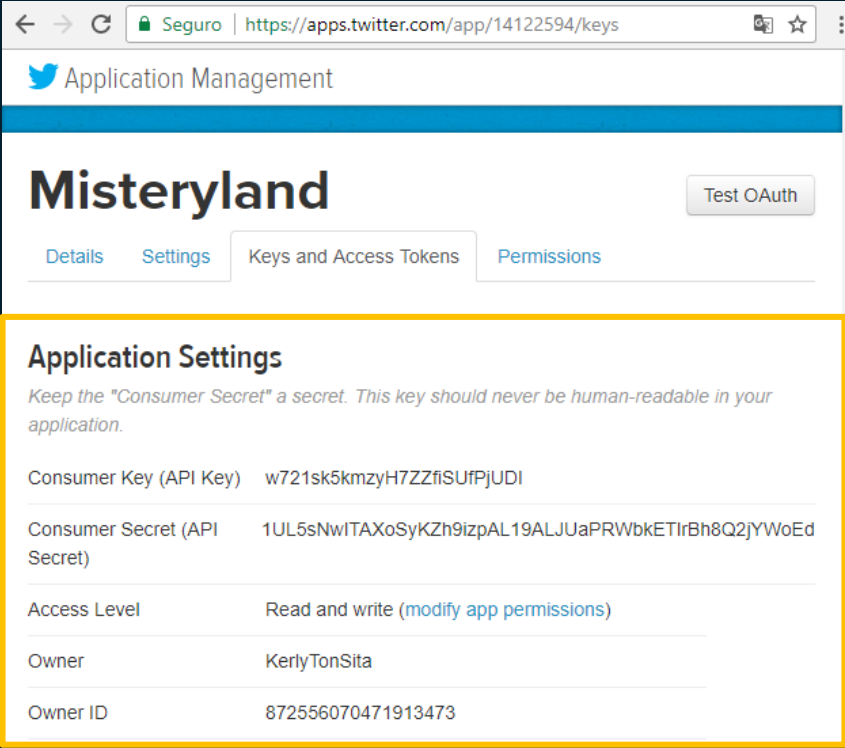


EXTRACCIÓN DE DATOS

La extracción de datos está destinada a mostrar cualquier correlación dentro del volumen de datos del sistema de información con el fin de detectar la tendencia de los datos.



EXTRACCIÓN DE DATOS



The screenshot shows the Twitter Application Management interface for an application named 'Mysteryland'. The browser address bar indicates the URL is <https://apps.twitter.com/app/14122594/keys>. The page has a blue header with the Twitter logo and 'Application Management'. Below this, the application name 'Mysteryland' is displayed in large black text, with a 'Test OAuth' button to its right. A navigation bar contains four tabs: 'Details', 'Settings' (which is active), 'Keys and Access Tokens', and 'Permissions'. The 'Settings' tab is highlighted with a blue underline. The main content area is titled 'Application Settings' and includes a warning: 'Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.' Below the warning is a table of application settings.

Consumer Key (API Key)	w721sk5kmzyH7ZZfiSUfPjUDI
Consumer Secret (API Secret)	1UL5sNwITAXoSyKZh9izpAL19ALJUaPRWbKETIrBh8Q2jYWoEd
Access Level	Read and write (modify app permissions)
Owner	KerlyTonSita
Owner ID	872556070471913473

EXTRACCIÓN DE DATOS

Name	Size	Number of Documents	Update Seq
_replicator	4.1 KB	1	1
_users	4.1 KB	1	1
ecuador	102.7 MB	15750	15750
mysteryland_k	1.0 GB	224379	224380

Showing 1-4 of [← Previous Page](#) | Rows per page: [Next Page →](#)
4 databases

EXTRACCIÓN DE DATOS

```
Seleccionar C:\Windows\system32\cmd.exe
Microsoft Windows [Versión 10.0.14393]
(c) 2016 Microsoft Corporation. Todos los derechos reservados.

C:\Users\usuario>C:\Users\usuario\Documents\BigData\curl-7.53.1\src
"C:\Users\usuario\Documents\BigData\curl-7.53.1\src" no se reconoce como un comando interno o externo,
programa o archivo por lotes ejecutable.

C:\Users\usuario>cd C:\Users\usuario\Documents\BigData\curl-7.53.1\src

C:\Users\usuario\Documents\BigData\curl-7.53.1\src>curl -X GET http://127.0.0.1:5984/mysteryland k/ al
l docs?include docs=true > datos.json
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total   Spent    Left   Speed
100 54.0M    0 54.0M    0     0  14.7M      0 --:--:--  0:00:03 --:--:-- 14.7M

C:\Users\usuario\Documents\BigData\curl-7.53.1\src>
```

EXTRACCIÓN DE DATOS

- Crear un archivo logstash.conf y agregar la siguiente configuración:

```
input{
  couchdb_changes{
    db=>"mysteryland_k"
  }
}
output{
  elasticsearch{
    host=>"127.0.0.1:9200" index => "mysteryland"
  }
}
```

EXTRACCIÓN DE DATOS

The screenshot shows the Kibana Overview page for Elasticsearch. The browser address bar displays `localhost:9000/#/overview?host=http%2F%2Flocalhost:9200%2F`. The page has a navigation bar with links for **overview**, **nodes**, **rest**, and **more**. A search bar at the top allows filtering nodes by name. The main content area displays a table of indices and a node named 'kerly'.

Indices Table:

Index Name	Shards	Docs	Size
clientes	5 * 2	48	90.90KB
index3	1 * 1	2	7.25KB
mysteryland	1 * 1	134,502	751.97MB

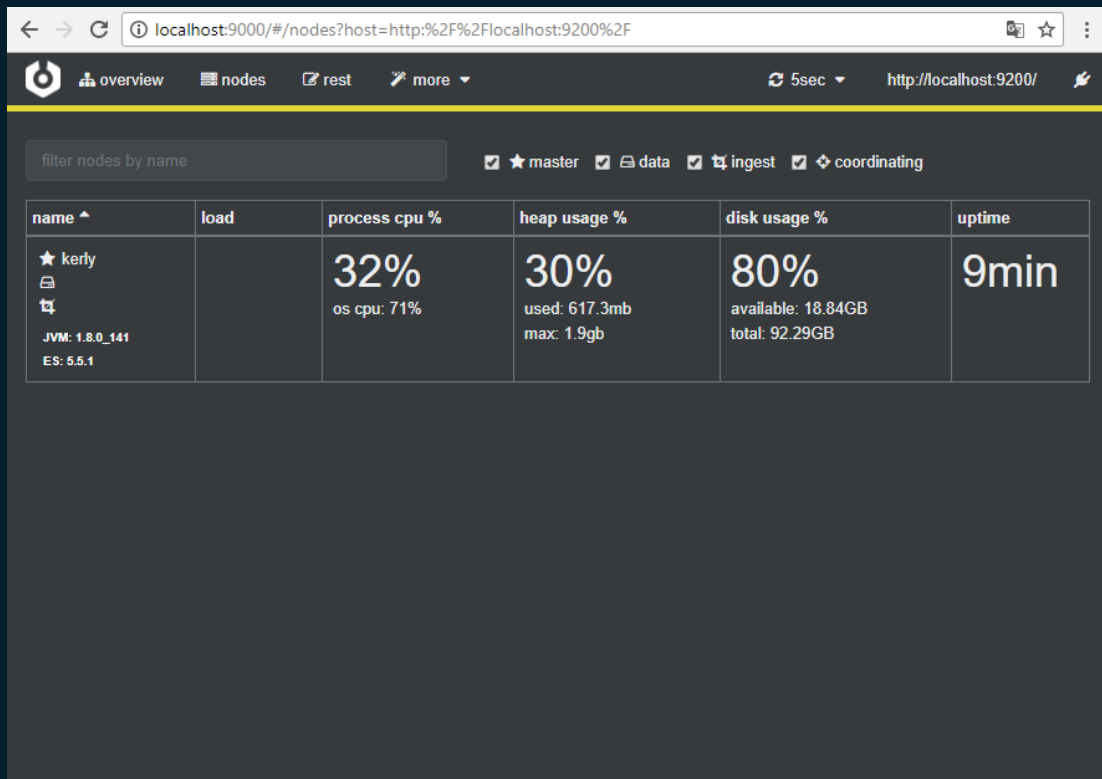
Node: kerly
IP: 192.168.0.107
Status: ● (Healthy)
Metrics: heap, disk, cpu, load

Shards: 11 unassigned shards (show only affected indices)

Shard Allocation:

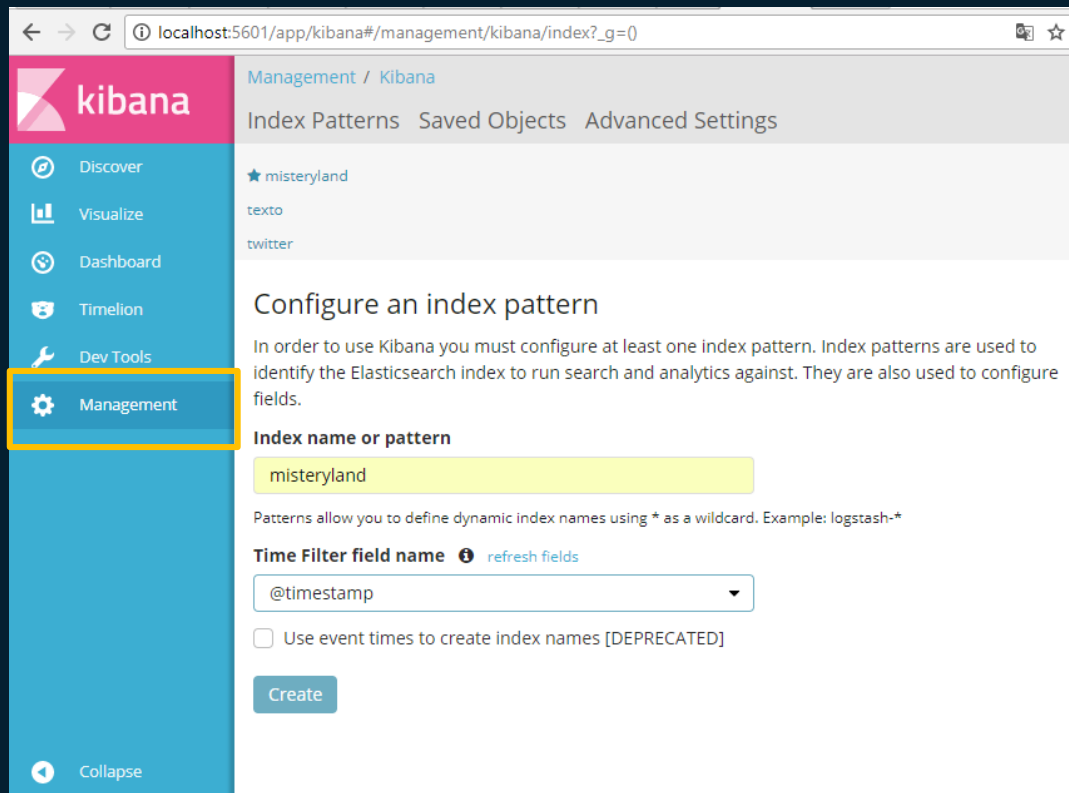
Index	Shard	Node
clientes	0	kerly
	1	kerly
	2	kerly
	3	kerly
	4	kerly
index3	0	kerly
	1	kerly
mysteryland	0	kerly
	1	kerly

EXTRACCIÓN DE DATOS



name ^	load	process cpu %	heap usage %	disk usage %	uptime
★ kerly 📄 🏷 JVM: 1.8.0_141 ES: 5.5.1		32% os cpu: 71%	30% used: 617.3mb max: 1.9gb	80% available: 18.84GB total: 92.29GB	9min

ANÁLISIS DE LA INFORMACIÓN



The screenshot displays the Kibana web interface in a browser window. The address bar shows the URL: `localhost:5601/app/kibana#/management/kibana/index?_g=0`. The left sidebar contains the Kibana logo and a menu with the following items: Discover, Visualize, Dashboard, Timelion, Dev Tools, and Management. The 'Management' item is highlighted with a yellow rectangular box. The main content area is titled 'Management / Kibana' and includes sub-links for 'Index Patterns', 'Saved Objects', and 'Advanced Settings'. Below these links, there is a star icon and the text 'mysteryland', followed by the words 'texto' and 'twitter'. The main heading is 'Configure an index pattern'. The text below explains that index patterns are used to identify Elasticsearch indices for search and analytics. A form field labeled 'Index name or pattern' contains the text 'mysteryland'. Below this, a note states: 'Patterns allow you to define dynamic index names using * as a wildcard. Example: logstash-*'. The 'Time Filter field name' is set to '@timestamp' in a dropdown menu, with a 'refresh fields' link next to it. There is an unchecked checkbox for 'Use event times to create index names [DEPRECATED]'. At the bottom of the form is a 'Create' button. The bottom of the sidebar has a 'Collapse' button.

← → ↻ ⓘ localhost:5601/app/kibana#/management/kibana/index?_g=0

kibana

Discover
Visualize
Dashboard
Timelion
Dev Tools
Management

Management / Kibana

Index Patterns Saved Objects Advanced Settings

★ mysteryland

texto
twitter

Configure an index pattern

In order to use Kibana you must configure at least one index pattern. Index patterns are used to identify the Elasticsearch index to run search and analytics against. They are also used to configure fields.

Index name or pattern

mysteryland

Patterns allow you to define dynamic index names using * as a wildcard. Example: logstash-*

Time Filter field name ⓘ [refresh fields](#)

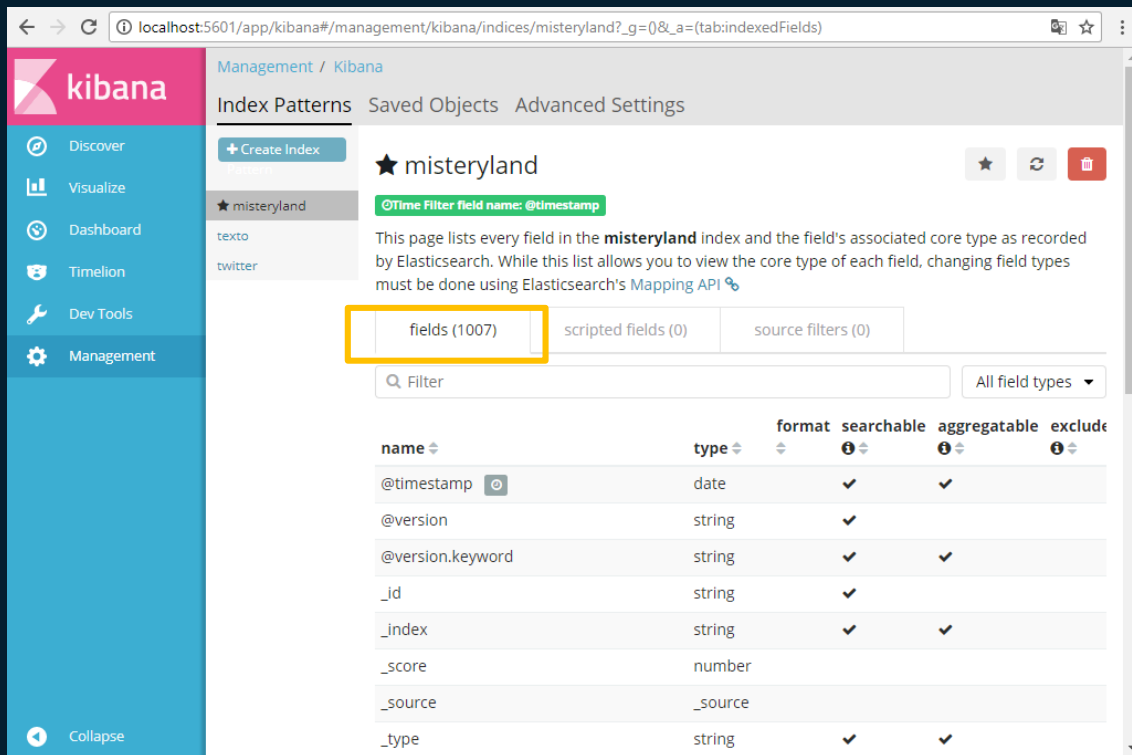
@timestamp

☐ Use event times to create index names [DEPRECATED]

Create

⌵ Collapse

ANÁLISIS DE LA INFORMACIÓN



Management / Kibana

Index Patterns Saved Objects Advanced Settings

+ Create Index

★ mysteryland

texto

twitter

Time Filter field name: @timestamp

This page lists every field in the **mysteryland** index and the field's associated core type as recorded by Elasticsearch. While this list allows you to view the core type of each field, changing field types must be done using Elasticsearch's [Mapping API](#).

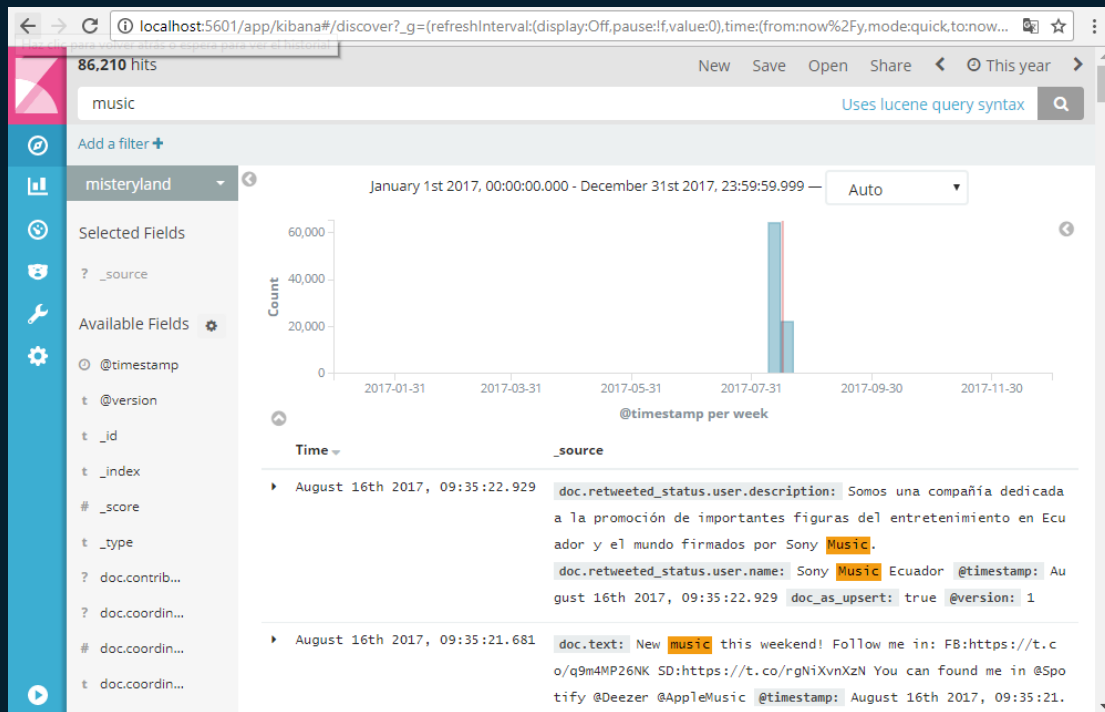
fields (1007) scripted fields (0) source filters (0)

Filter

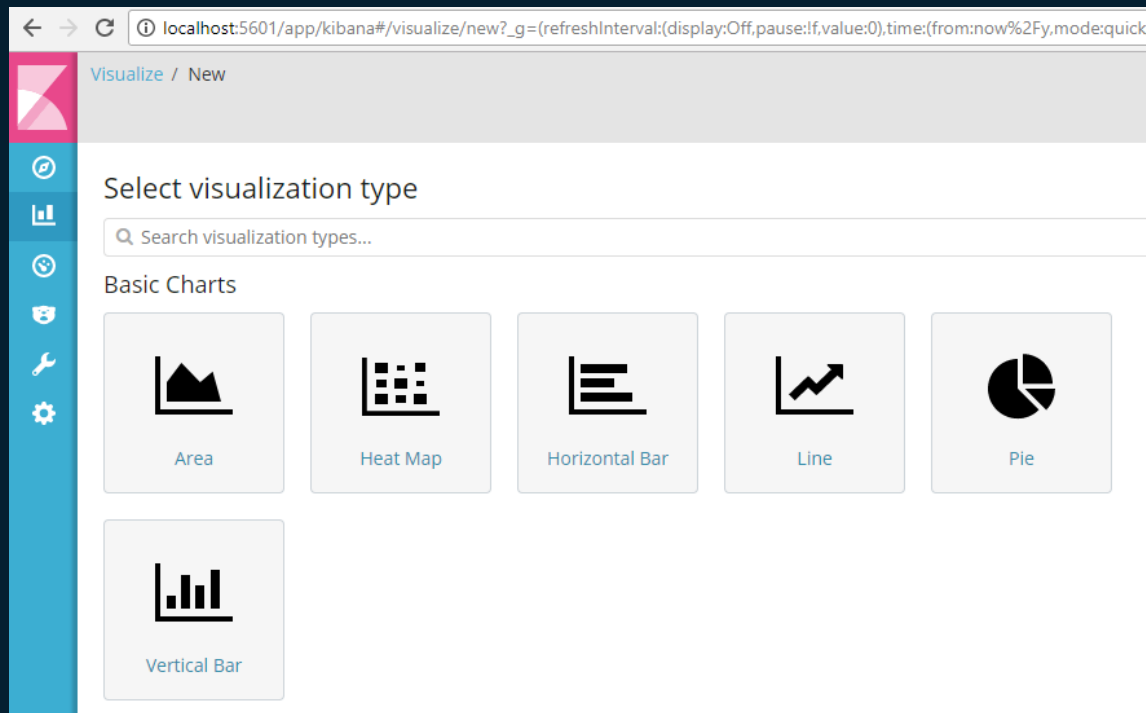
All field types

name	type	format	searchable	aggregatable	exclude
@timestamp	date		✓	✓	
@version	string		✓		
@version.keyword	string		✓	✓	
_id	string		✓		
_index	string		✓	✓	
_score	number				
_source	_source				
_type	string		✓	✓	

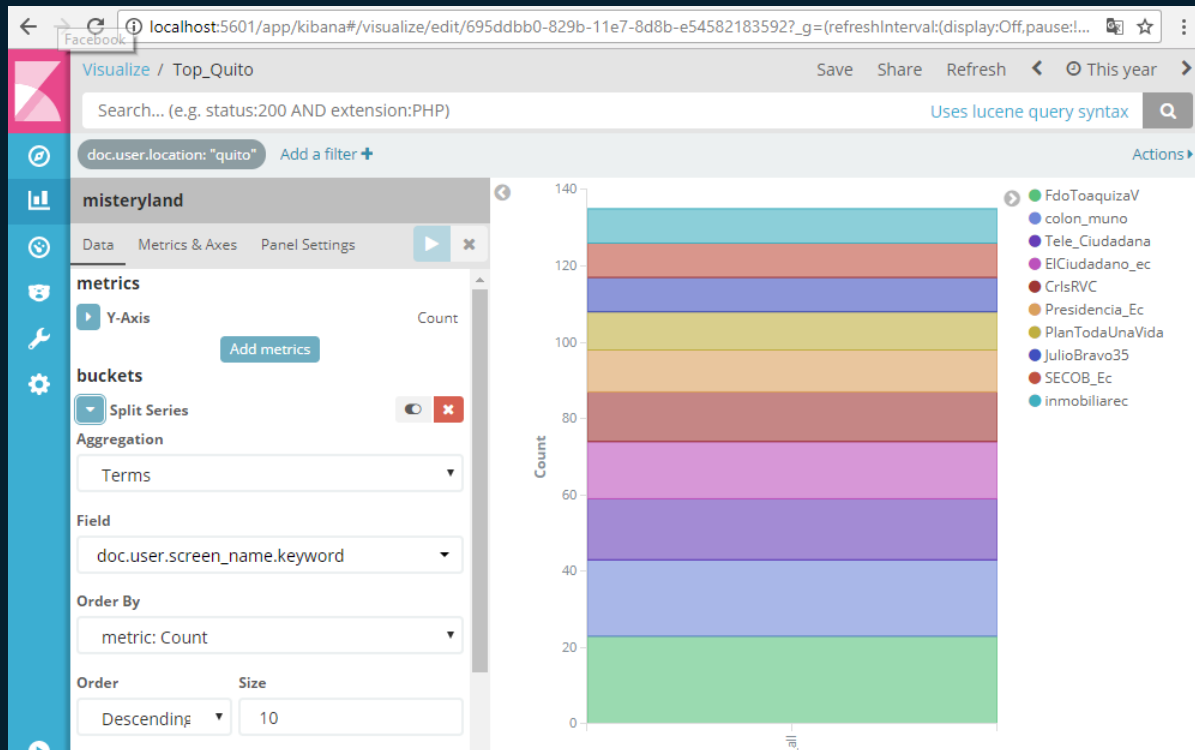
ANÁLISIS DE LA INFORMACIÓN



ANÁLISIS DE LA INFORMACIÓN



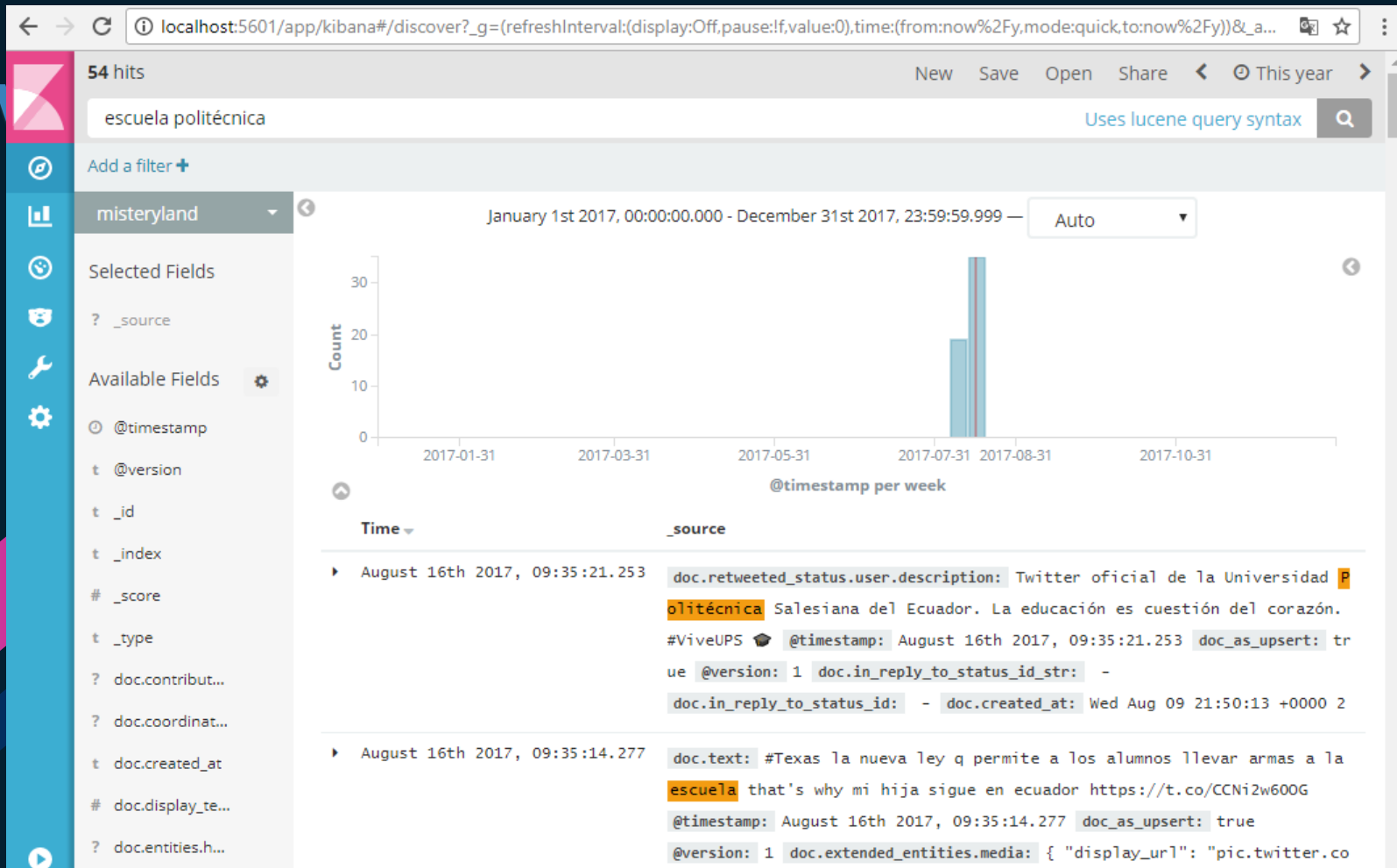
ANÁLISIS DE LA INFORMACIÓN



RESULTADOS OBTENIDOS

Caso de estudio 1

En la búsqueda de la EPN en las redes sociales no se pudo obtener datos relacionados y podemos concluir que la EPN no tiene mucha presencia en las redes sociales relacionado al evento Mysteryland



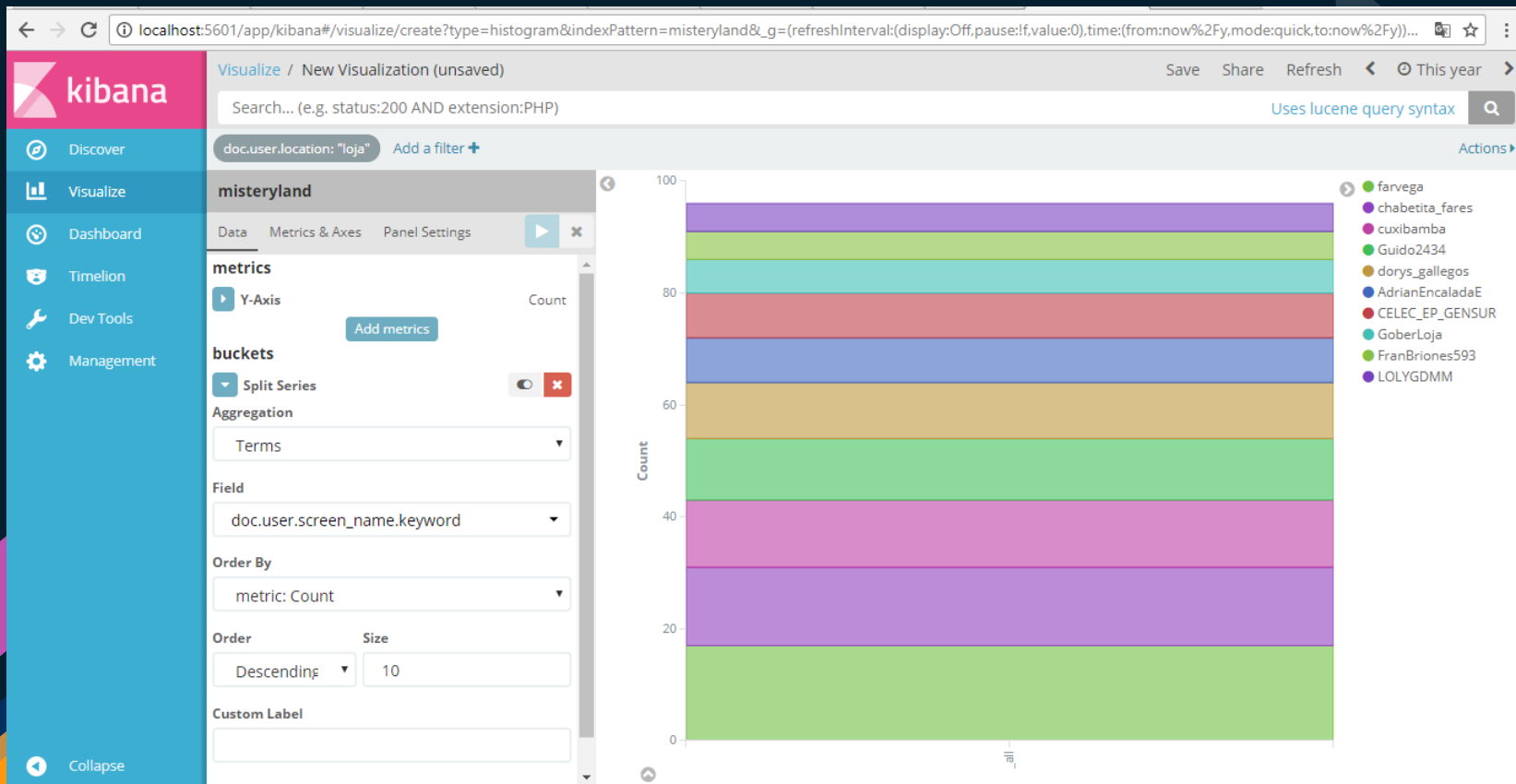
RESULTADOS OBTENIDOS

Caso de estudio 2

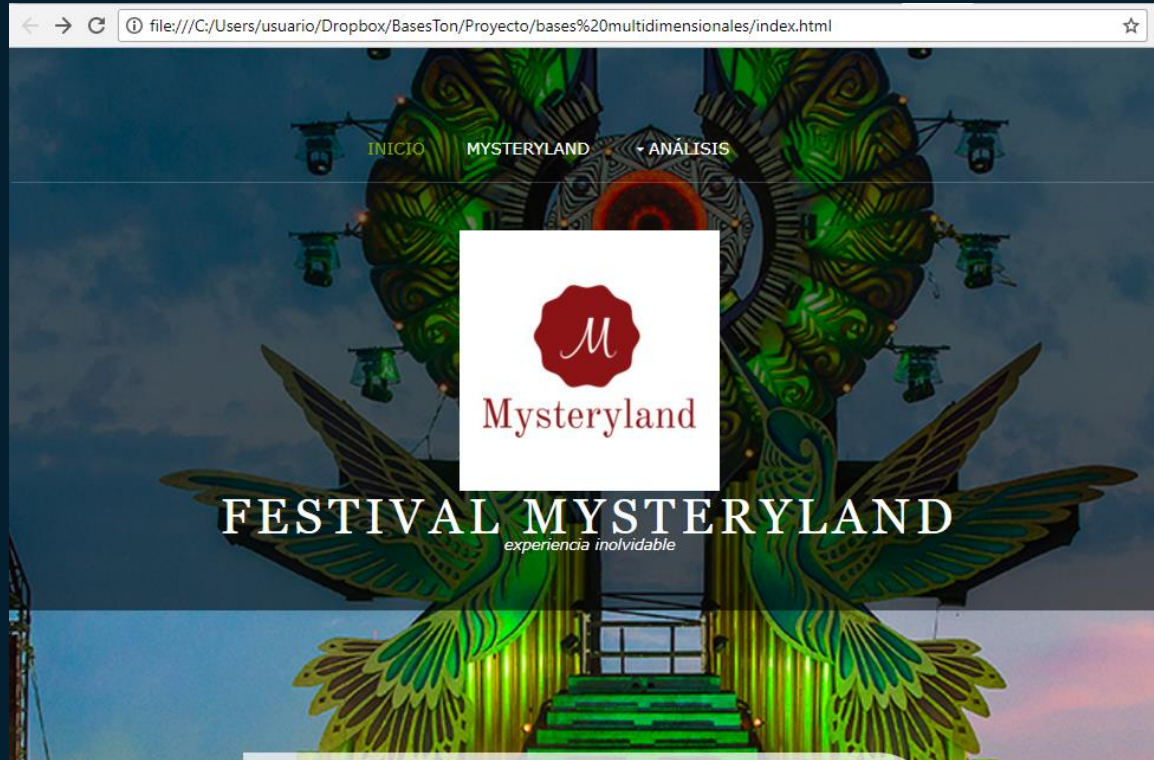
En este caso se analizan el top 10 de los twitteros de cinco ciudades de Ecuador.

TOP 10 LOJA

En el top 10 de los twiteros en la ciudad de Loja se encontró al mayor twitterero “Farvega” con 17 tweets en menos de un mes. Según el análisis realizado Farvega es una persona que tuitea noticias acerca de las actividades de gobierno.



VISUALIZACIÓN DE LA INFORMACIÓN



CONCLUSIONES

- ❖ Existen eventos que contratan publicidad en redes sociales, los cuales realizan marketing continuamente, la promoción se realiza cada segundo, estos datos entorpecen el análisis ya que para el análisis únicamente se necesita datos que fueron realizados por personas porque si no tendríamos un análisis incorrecto y nos alejaríamos del objetivo del análisis.
- ❖ Al recolectar los datos, pudimos contar con un número elevado de tweets de las personas que realizaron un hashtag a la página de Mysteryland, y por los cuales pudimos realizar un Mapreduce y con ello logramos determinar el país que más tweets realizaba al día. Además por medio de Kibana se obtuvo una visualización del usuario que constantemente realiza hashtag.
- ❖ La indexación de la base con la herramienta elasticsearch es el proceso previo para la utilización del LMNZ Cerebro, en este punto se recomienda realizar una indexación de datos con un mapping de los datos, para no tener problemas con los parámetros.

RECOMENDACIONES

- ❖ Se recomienda que al momento de enlazar la base de datos CouchDB con la herramienta elasticsearch, se utilice un mapping para la información, aunque esto no es obligatorio, se lo debe hacer para que todos los datos se han consistentes y no tener datos que tengan parámetros sin llenado, con el mapping se le asignara un valor por default según se ha el caso.
- ❖ Para trabajar con las herramientas mencionada en este informe, se debe de tener en cuenta las versiones de cada herramienta, ya que en algunas herramientas no son compatibles con algunas versiones. De las herramientas.
- ❖ Se recomienda no realizar el análisis con la información que tiene como origen "publicidad", ya que tendríamos un análisis alejado del objetivo.

Gracias



Realizado por :
Katherine Lasluisa
Kerly Naranjo
Ricardo Sinchiguano